Time-Frequency Distribution of Music based on Sparse Wavelet Packet Representations

Endelt, Line Ørtoft

Publication date:
2005

Document Version
Early version, also known as pre-print

Time-Frequency Distributions of Music based on Sparse Wavelet Packet Representations

Line Ørtoft Endelt and Anders la Cour-Harbo
Aalborg University
Department of Control Engineering
Frb. Vej 7C, 9220 Aalborg East, Denmark
{oertoft, alc}@control.aau.dk

Abstract

We introduce a new method for generating time-frequency distributions, which is particularly useful for the analysis of music signals. The method presented here is based on ℓ^1 sparse representations of music signals in a redundant wavelet packet dictionary. The representations are found using the minimization methods basis pursuit and best orthogonal basis. Visualizations of the time-frequency distribution are constructed based on a simplified energy distribution in the wavelet packet decomposition. The time-frequency distributions emphasizes structured musical content, including non-stationary content, by masking the energy from less structured music instruments. We present four examples for visualizing structured content, including vocal and single instrument.

Keywords: wavelet packet, redundant representations, time-frequency distribution, music.

1 Introduction

The basic idea of redundant representation is that by employing a richer dictionary there is potential for faster decay of the weight coefficients in the representation than is the case for a 'sufficient' dictionary. This seemingly appealing fact can then be explored for the purpose of compression, denoising, feature extraction, and other applications. However, the use of redundant representations holds many challenges compared to the non-redundant case, and consequently redundant representations as a mean to achieve sparseness in signal representation has been investigated by many researchers in recent years. Further, these challenges often inhibits the use of redundant representations in applications, which otherwise would seem to benefit from the redundancy. One example of an application where non-redundant representations is very common is time-frequency distributions.[1]

In this work we are interested in using redundant representations of music signals for extracting or emphasizing certain time-frequency (TF) related features. In particular, we demonstrate how redundant wavelet representations can be used for generating 'scalograms' with specific properties not usually found in standard wavelet representations. We approach this from a mathematical point-of-view rather than applicational, and consequently our work is more on what can be achieved using this specific method rather than attempting to address a particular task, like detection of fundamental frequency, onset detection, beat estimation, and so on. This makes our approach somewhat different from other works in this field, see for instance [23, 19, 22, 5, 20].

One of the features that comes out of using redundant representation on music signals is a visualization of the presence of a distinct melody from a human vocal or a single instrument.

2 Methodology

When applying a Fourier or Wavelet transform to describe a music signal a complete description is achieved, which is usually more sparse than the original signal, and

[1] We disregard sliding windows, over-sampled FFTs, and the like as truly redundant representations.

often more meaningful in terms of what the music signals contains. But although the description is complete it is not necessarily useful or sufficiently sparse. The main reason being that a music signal typically is a very complex signal consisting of a variety of different events that does not fit a single type of dictionary, such as pure frequencies (Fourier or sin/cosine). Ideally, we can think of music as being well represented in a 'music instrument' domain, i.e. a domain where each instrument is somehow represented in the dictionary. The representation of the individual instruments is then composed of a number of frequency and time localized events, and each of these are in turn associated with a set of parameters, like attack, sustain, decay, vibrato, intonation, and so on. Adding to the complexity is the fact that in some cases rather different settings of the above leads to sounds almost indistinguishably by human, while in other cases even small alternations can lead to very detectable changes in the sound. As an example the phase of frequencies are not detectable (directly) by humans, while small changes in pitch are very detectable.

While it is relatively simple to bring music from a description in the above form (MIDI is an example of such a type of representation) to a digital waveform of decent quality, it is at present in general not possible to go the other way, to do automatic scoring. The mathematics of signal representation is still far from being able to handle this complexity.

However, we are indeed able to push the limit further than the standard dictionaries by employing the redundant representations. They are one step closer to the ideal representation in the sense that the redundancy allows much more freedom in the choice of dictionary content, the so-called atoms. This freedom comes specifically with redundant representations; while orthogonality introduces several very nice properties in the signal representation, it simultaneously introduces (unnecessarily) severe restrictions. This restriction is clearly seen in any orthogonal TF distribution (TFD) like a spectrogram or scalogram of a sinusoid with a single spike, where the frequency event and time event cannot be simultaneously well localized.

Although the dictionaries in this work are redundant they are still composed of 'standard' dictionaries. In particular, we have chosen to focus on a redundant wavelet dictionary. The reason for not including a 'pure tone' dictionary is mainly that the frequency localization property of the Fourier dictionary is to some extend present in a multi-level wavelet packet (WP) transform. However, ongoing work by the authors include TFD based on combined Fourier and WP dictionaries. We denote the TFD based on a redundant representation RR-TFD.

The following sections will describe the details of the redundant signal representations, the WP dictionaries and two minimization methods for finding sparse representations of signals in dictionaries. There is a description of the calculation setup, and how the RR-TFD shown in section 3 are produced from the energy of the coefficients of a redundant representation of a music piece.

2.1 Redundant Signal Representations

A dictionary for $\boldsymbol{R}^N$ is a set of M vectors in $\boldsymbol{R}^N$, $M \geq N$, that span $\boldsymbol{R}^N$. The elements of a dictionary are called atoms. Thus, for any signal $\boldsymbol{b} \in \boldsymbol{R}^N$ we can find $\boldsymbol{x} \in \boldsymbol{R}^M$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}$ is an $N \times M$ matrix with the atoms as columns. Since this set of equations is overcomplete it has infinitely many solutions, and we therefore need to specify, which solution is desired. In the present setting we want a sparse solution, and we choose to use the ℓ_p norm, $p \leq 2$, although many other measure could be used.

Consequently, the initial problem of interest here is

$$\min \|\boldsymbol{x}\|_p \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

where $\boldsymbol{b}$ is the music signal, $\boldsymbol{A}$ is the dictionary, and $\boldsymbol{x}$ is the representation. When the dictionary contains more than N elements this representation is not unique, and this is the source of the increased flexibility in choice of representation compared to the orthogonal case.

There are many methods for constructing a sensible dictionary, as examples see e.g. [2, 1, 14]. It can contain harmonic and Gabor waveforms, wavelets, chirps, spikes, and so on. It can also contain elements learned by training on a set of similar signals. In this work we use a WP dictionary (see section 2.2).

There exists a number of optimization methods for finding the most sparse representations in a dictionary for a given signal, i.e. for solving (1). We have chosen the two rather different methods basis pursuit and best orthogonal basis. This is further described in section 2.3.

2.2 Wavelet Packets

A WP transform is the application of a pair of wavelet FIR filters combined with a hierarchical way of applying the filters. The filters are a low pass filter $h[n]$ and a high pass filter $g[n]$, where the filter taps (impulse response) satisfies a series of constraints [25]. In general, the WP is applied the following fashion. First, the filters are applied on the original signal $\boldsymbol{b}$ of length N, followed by a down sampling resulting in two new signals (or filter outputs), $\boldsymbol{x}_{\text{low}}$ and $\boldsymbol{x}_{\text{high}}$, with

$$x_{\text{low}}[k] = \sum_n b[n]h[2k-n] \tag{2}$$

$$x_{\text{high}}[k] = \sum_n b[n]g[2k-n]\ . \tag{3}$$

Next, the low and high pass filters are applied on both $\boldsymbol{x}_{\text{low}}$ and $\boldsymbol{x}_{\text{high}}$. This procedure is continued until the desired frequency resolution is reached. Figure 1 shows the hierarchical structure of the outcome of the WP. The first level contains the original signal, the second level contains the output of the low and high pass filters (when applied to the original signal). This is continued, so that below each box in the scale frequency diagram is the outcome, boxes, of applying the low pass filter (on the left) and the high pass filter (on the right) on the signal in that particular box. Each level in Figure 1 contains N elements, and since the

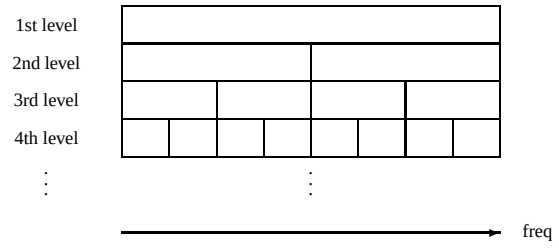

Figure 1. Four levels of a WP transform, each box corresponds to a filter output.

filters are invertible, the signal can be reconstructed from the information in one level. Coefficient j at the ith level, α_{ij}, in the WP decomposition corresponds to one particular waveform $\boldsymbol{\phi}_{ij}$ in the time domain. This can be exploited to generate the transform matrix, and thus the atoms in the dictionary. In short, this is done by assuming $\alpha_{ij} = 1$ and all other coefficients are 0, and then apply the inverse WP transform. See [17] for a more thorough description of this construction and of the generation of the transform matrix.

Therefore the output of the WP transform is the coefficients

$$\alpha_{ij} = \sum_{n=1}^{N} \phi_{ij}[n] b[n] \ . \tag{4}$$

The collection of the waveforms $\boldsymbol{\phi}_{ij}$ is a dictionary for $\boldsymbol{R}^N$. So the WP dictionary is basically the matrix embodiment of the linear WP transform.

Using the notation from (1) each waveform $\boldsymbol{\phi}_{ij}$ corresponds to a column in $\boldsymbol{A}$. Note that in any software coding for the purpose of actually computing the coefficient vector $\boldsymbol{x}$ the filter implementation is used rather than the matrix implementation, as the former is an $O(N \log N)$ and the latter is $O(N^2)$.

The representation of a signal in a WP dictionary is not unique; there are infinitely many solutions to the overcomplete system of equations in (1). Consequently, some methods for choosing the optimal solution is needed. In particular, we want sparse representation. Optimal and near-optimal sparse solutions can be achieve in several ways. This is the subject of the following sections.

2.3 Finding the Sparsest Representation

The perhaps most obvious choice for the sparseness measure in (1) is the ℓ_0 norm since this measures the number of non-zero entries. Unfortunately, finding this particular representation is in general NP hard, and thus not feasible for even moderately sized problems. Also, the probability that an arbitrary signal in $\boldsymbol{R}^N$ lies in any of the finitely many $N-1$ dimensional subspace spanned by the atoms is 0, so in (1) the probability of $\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 = N$ is 1. However, there is a series of results on the relations between solutions to (1) for varying p. In particular, it has been shown [12, 6] that under some separability conditions imposed on the dictionary (basically requiring sufficiently different atoms) the solution to (1) is the same for $p = 0$ and $p = 1$. Similar results exists for more general sparseness measures [11]. Although these conditions are rather difficult to fulfill, in particular when constructing dictionaries by combining various collections of waveforms, it indicates that the ℓ_1 norm is a feasible spareness measure. In the following only this norm is considered. For a discussion of ℓ_0 versus ℓ_1 and the use of ℓ_1 as sparseness measure, see for instance [2, 12, 6].

Except for $p = 2$ no known method exists for analytically determining the solution to (1). In the $p = 2$ case the solution is given by the the pseudo inverse (also called Moore Penrose inverse, method of frames [3]), but this solution is in general not sparse as no entries in $\boldsymbol{x}$ are vanishing. For $p = 1$ it is a challenge to actually find the minimizer $\boldsymbol{x}$, and the challenge varies (occasionally a lot) with the choice of method, dictionary, measure, and signals.

Fortunately, there do exists general methods for iteratively approximating the solution $\boldsymbol{x}$ when $p = 1$ (and for other similar sparseness measures). Some examples are linear programming [2] (also known as basis pursuit), quadratic programming [9], minimum fuel neural networks [24, 18], and FOCUSS [10] (actually solves for some $p < 1$). Sub-optimal solutions can be obtained by various types of matching pursuit [16, 21], alternating projections, and best orthogonal basis [25, 13, 17].

In this paper we use basis pursuit (BP) and best orthogonal basis (BOB). The overall advantage of the best orthogonal basis search is that it is much faster than basis pursuit (for a comparison of computation times, see [8]).

In our experience the basis pursuit method is capable of producing a truly sparse representation of a music signal. That is, not only is the ℓ_1 minimal, but the decay of the entries in $\boldsymbol{x}$ is also satisfying. In particular, when using a WP dictionary with the appropriate mother wavelet, or indeed a combined WP and cosine packet (CP) dictionary. For an example of the latter see [15], where a representation of a piece of music is found using basis pursuit on a large dictionary consisting of five sub-dictionaries. The representation

has a good resolution, where the signal is divided into a 'beat/drum' part (described in a WP) and a 'melody' part (described in a CP).

A best orthogonal basis search is an adaptive way to choose an orthogonal subset of the atoms in dictionary with a particular structure, such as found in the WP or CP dictionaries, so that the coefficients of the representation are minimized according to some cost function (like ℓ_p norm or entropy). This is a fast way to find a fairly sparse representation, but it is restricted by the orthogonality constraint (as discussed earlier), which in turn restricts the possibly tiling of the TF plane.

2.4 Calculation the Sparse Representations

We want to find the sparsest representation of a collection of music pieces for the purpose of making RR-TFD of each piece. This section contains a description of the calculation setup.

Basis pursuit and best orthogonal basis are applied for finding sparse representations of music signals in a WP dictionary with 9 levels generated with the least asymmetric (almost linear phase) Daubechies wavelet of filter length 12 (also known as Symlets, see [4]).The choice of wavelet is based on the considerations in [7]. BOB is applied with the ℓ_1 norm as cost function. Representations are found on a collection of approximately 400 music signals from various genres. The sample frequency is 44.1 kHz, and the sampling starts 60 seconds after the beginning of the song and lasts 30 seconds. Each of these music sequences are divided into non-overlapping windows of length $L = 8192$ samples, so each $\boldsymbol{b}$ is an 8192×1 vector, and basis pursuit is applied on each of the $\lfloor 30 \times 44100/8192 \rfloor = 161$ analysis windows. Thus, the are 161 representations found for each music piece.

The basis pursuit and best orthogonal basis implementations described in [2] are used for all computations of representations shown in this paper.

The calculations are obtained as part of a larger calculation setup which is described in [8]. In this setup five different dictionaries, five different minimization methods, and four different window lengths are combined, with and without down sampling, giving 168 calculation combinations, which are all applied on approximately 400 pieces of music. Since storing all the representations found for this calculation setup required too much storage space (in the order of terabytes) a number of measures applied to the coefficients has been stored instead. The RR-TFDs described in Section 2.5 are based on the ℓ_2^2 measure applied to the boxes in the WP decomposition, see Figure 1, that is, the energy distribution of the coefficients of the representations found of music signals in a WP dictionary. This is explained in more detail in the following section.

2.5 Time-Frequency Distributions

The coefficient vector $\boldsymbol{x}$ of the representation of an analysis window $\boldsymbol{b}$ is split up according to the hierarchical structure in the WP (see Figure 1) such that each vector contains the coefficients of the atoms corresponding to one particular filter output (or box). The notation used for the resulting coefficient vectors is shown in Figure 2.

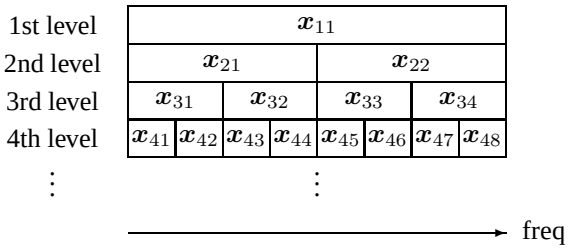

Figure 2. The coefficient vector $\boldsymbol{x}$ of a representation in a WP dictionary divided according to the WP hierarchy.

The number of elements in the coefficient vector $\boldsymbol{x}_{ij}$ is $\#\boldsymbol{x}_{ij} = L/2^{i-1}$ and the energy in $\boldsymbol{x}_{ij}$ is $\|\boldsymbol{x}_{ij}\|_2^2$. These energies are the measurements available from the calculations described in the previous section.

The standard approach for making a TFD of a WP decomposition is to find the best basis (i.e. choosing the best boxes in the decomposition) and then map each coefficient in the chosen basis to a single tile. Due to the interpretation of the WP atoms this particular approach produces a disjoint tiling covering the entire TF plane.

Our approach for constructing a RR-TFD is slightly different. We do not use the individual coefficients in the sparse representation. Instead we use the total energy of each box in the decomposition in the following way. Define 'frequency' vectors

$$\boldsymbol{f}_{ij} = \mathbf{1}_{2^{9-i}\times 1} \otimes \frac{\|\boldsymbol{x}_{ij}\|_2^2}{2^{9-i}}, \quad i = 1, \ldots, 9 \,,$$

with $\otimes$ being the matrix direct product. Note that there are 9 levels in our WP dictionary. This produces vectors that has the same length as the original coefficient vectors $\boldsymbol{x}_{ij}$, but with the entries equal to the average energy of the coefficients in the box. Due to the frequency localization property of iterated wavelet filters and Parselval's equality this gives a simplified version of the frequency energy distribution in the redundant representation.

To generate the TFD define

$$\boldsymbol{f} = \sum_{i=1}^{9} \begin{bmatrix} \boldsymbol{f}_{i1} \\ \boldsymbol{f}_{i2} \\ \vdots \\ \boldsymbol{f}_{i2^{i-1}} \end{bmatrix}_{256\times 1}$$

and plot $\boldsymbol{f}$ as a column in the TFD at the time instant corresponding to the position of the window in the original 30 second signal. Plotting all 161 $\boldsymbol{f}$ vectors produce an RR-TFD plot. All plots in this paper are visualized with in dB scale. Note that the frequency vector $\boldsymbol{f}$ satisfies $||\boldsymbol{f}||_2^2 = ||\boldsymbol{x}||_2^2$.

This procedure results in a tiling where all tiles have equal width and varying height, varying according to frequency content. Consequently, the tiles are overlapping, and thus each point in the RR-TFD contains information from all levels in the WP decomposition.

The RR-TFD is constructed such that elements that have the same behavior (structure or frequency) over an entire window will be seen more clearly in the mapping, since these elements are described at the lower levels of the WP, and therefore the energy is only distributed in a few of the frequency bins. This means that time localized musical content has a tendency to be less visible in the TF plane, and thus we expect the harmonic content to be emphasized, even when this is not stationary. This is an important feature for this methods since it is not directly available with standard TFD based on harmonic dictionaries.

Note that in terms of mathematical properties the RR-TFD has in general the same properties as the scalogram (i.e. positivity, marginals not satisfied, etc.). Note also that the distributions plotted in the following sections are only the lowermost one fourth of the frequency interval, since the higher frequency part only contains vertical lines, that extends from the lower frequencies.

3 Results

To demonstrate the RR-TFD based on a wavelet dictionary this section shows the method applied to four different music excerpts:

1. Celine Dion “River Deep Mountain High”, soft rock with female vocal, drums, guitar, synthesizer,
2. DAD “Candid”, with male vocal, heavy drums, and electrical guitar.
3. The Sandmen “Don’t let me down”, primarily drums, bass, and male vocal.
4. An orchestral piece with a predominant oboe.

3.1 Vocal

The first example is shown in Figure 3. Three distributions are shown, the first two plots are RR-TFD using BOB and BP, respectively, and the third plot is a standard spectrogram made by an STFT. The TF resolution and the window length and shape are the same in all three plots.

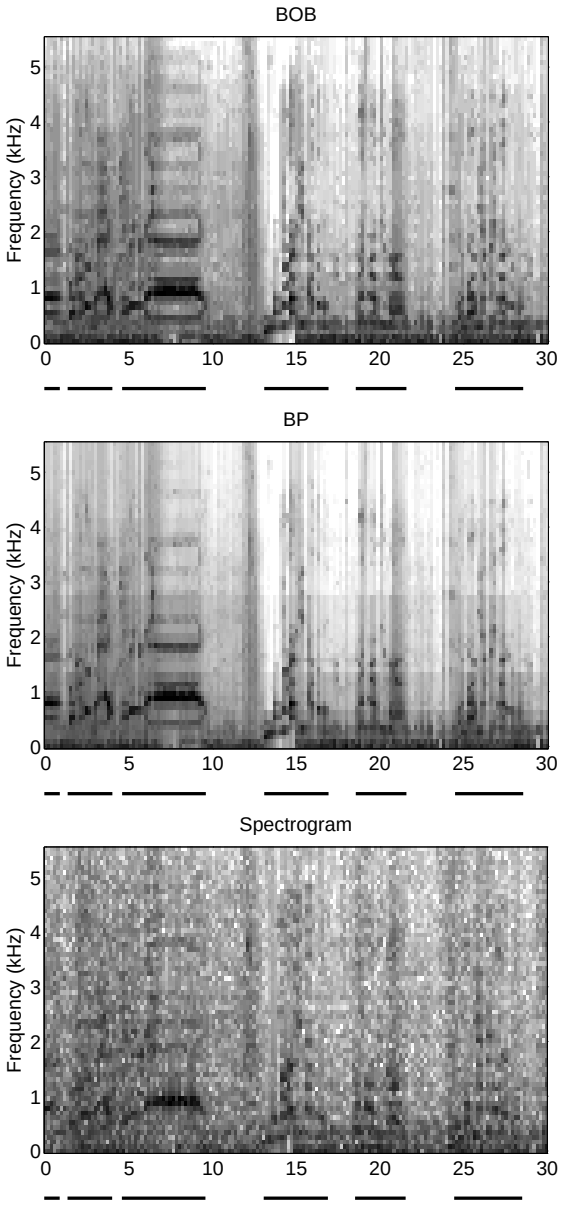

Figure 3. TFD of ‘River deep mountain high’ by Celine Dion. The two upper plots are RR-TFD based on BOB and BP, respectively. The third plot is a spectrogram. The lines under the plots mark when there is a vocal present in the music (determined manually by the authors).

This song contains a distinct female voice mixed with a few instruments. While the voice is very clearly heard in the music for a human listener, it is only just visible in the spectrogram. This is because the music instruments contain energy at most TF tiles (as discussed previously), and consequently, the spectrogram to some extent resembles that of noisy sound. In comparison, the RR-TFD presents a somewhat less noisy, and thus more clear, image of the predominant structure in the music. In particular, the BP based

RR-TFD shows a more clean TFD.

It is importance to note that the RR-TFD enhances the musical structure by 'disregarding' the less structured parts. This is seen in this example by the fact that the structure found in the BP RR-TFD is also present in the spectrogram; it is just less visible because the other musical content tends to disguise it.

The BP RR-TFD and the BOB RR-TFD are very much alike, but the structures in the signal is seen a bit more clearly in the BP RR-TFD. Therefore only the BP RR-TFD is shown for the following three examples.

The second example is the song "Candid" by DAD, a heavy rock music piece with male vocal and significant drums and guitar. Figure 4 shows the BP RR-TFD and the spectrogram. While the human vocal is still occasional visible in the BP RR-TFD it is much less significant than in the previous example. However, the spectrogram is close to useless in terms of visualization of structure. Note that

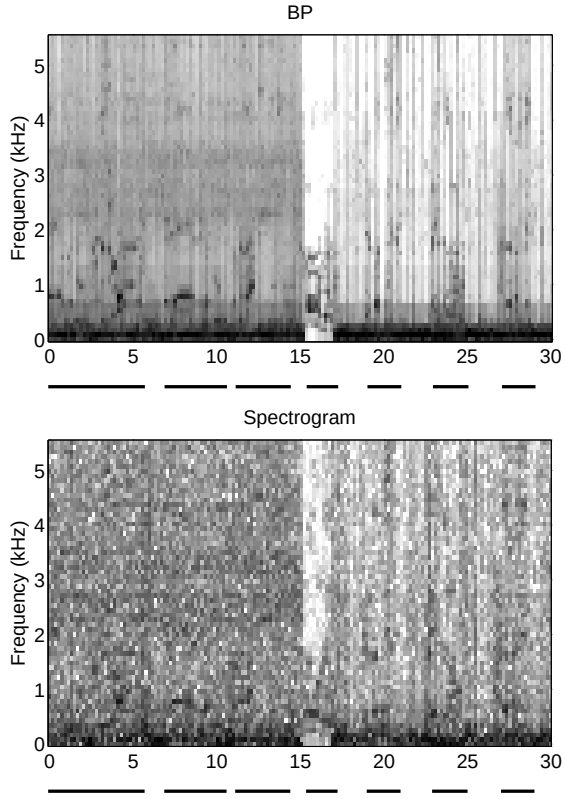

Figure 4. TFD of "Candid" by DAD.

in the entire excerpt, except the 2 seconds in the middle contains heavy drums. The first half part of the signal also contains electrical guitar. It is possible in the RR-TFD to see a slightly increased energy in the 500-2000 Hz region for the vocal parts. In the last half part of the excerpt there is a larger correlation between the vocal parts in excerpt and the "increased energy" areas, than in the first half. So it seems that the electrical guitar, which is very pronounced in the signal, do not get masked.

The third example is the song "Dont let me down", by the danish band "The Sandmen". It is a semi-hard rock music piece, similar to the previous example, but with less significant music instruments. In particular, the drum is less pronounced, but still fairly strong. Figure 5 shows the BP RR-TFD. The presence of a vocal is clearly seen, and the drums are efficiently suppressed. At the beginning of the vocal sequences, the boundary between the non-vocal and the vocal parts is seen very clearly.

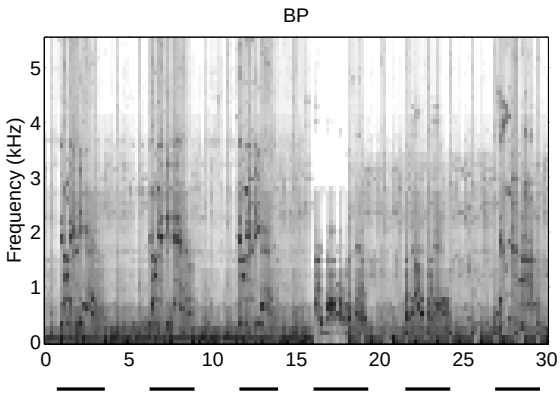

Figure 5. RR-TFD of "Dont let me down" by The Sandmen.

3.2 One Instrument

While the previous three examples emphasized the vocal as the predominant structure, this last example shows how an instrument can generate the main structure. In Figure 6 is a RR-TFD of a 30 second excerpt from an Oboe concerto. There are a few other instruments playing in this excerpt besides the oboe. The immediate impression from the RR-TFD is that not only is the structure clearly present. The actual notes can (almost) be determined from the plot. For comparison, the notes are also shown in the figure.

4 Discussion

We have present a time-frequency distribution based on the energy distribution of a sparse representation found in a redundant WP dictionary. The purpose of this TFD is to visualize main structures in music piece, in particular to be able to extract such structure in the presence of several other music instruments with highly varying TF content.

The method used here to visualize the energy distribution is thus able to mask instruments such as a drum, which

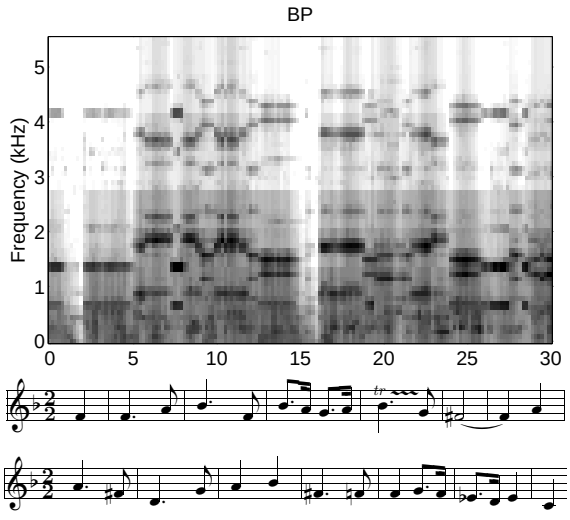

Figure 6. TFD of an excerpts from Händels Oboe Concerto no. 3 in G minor. Below is the score for this 30 second excerpt. The "line break" in the score occurs at 18 seconds. Notice that the duration of each note in the score is determined by the note-type, not the distance to the following note.

in standard harmonic dictionaries tends to produce noise-like TF structure when analyzed with long windows. This is clear from the spectrograms in Figure 3 and 4. This effect in turn makes structured sound such as human singing voice more visually apparent. The effect is greater for a female voice, since the frequencies, most often, are higher than for a male voice.

In general, whenever there is a distinct vocal it shows up in the mapping, even when the signal contains drumming, and the singing voice and the drum beats perceptually are equally strong. This tendency do get disturbed if the vocal is a very deep male voice, or if the music piece contains electrical guitar.

When the signal contains an instrument playing one tone at a time, it is almost possible to "follow" the notes in the TF plot.

In the current presentation[1] we have relied solely on wavelet dictionaries. There are two reasons for this choice. Firstly, the wavelet transform has well-known and fairly simple interpretation in terms of time-frequency, and secondly, any wavelet dictionary has an $N \log N$ implementation making it significantly faster than a matrix multiplication. While these two properties are arguably very useful in music analysis due to the time-frequency content and the length of music signals, the nature of our investigation easily leads to the question of learned dictionaries. That is, would it be advantageous to use dictionaries based on properties inherited directly from the given signal class rather than dictionaries that just happens to have fairly appropriate properties?

Since we want to use the redundant representations for classification purposes it seems obvious that learned dictionaries should perform better than any 'standard' dictionary, at least for the signals in the set used for generating the dictionaries, and possibly for other very similar signals. This is to be understood in the sense that there exist a representation, which for a given sparsity measure and signal set produces the optimal separation of the signals in coefficient space. If the representation is restricted to an orthogonal basis (in which case there is no redundancy) the Karhunen-Loeve transform would be optimal.

So data learned dictionaries (see e.g. [1, 14]) are interesting to investigate in relation to feature extraction of music. Ideally, such learned dictionaries might lead to an 'instrument transform' and/or a 'vocal' transform. But using a data learned dictionary might also lead to an increased computation time, since the new dictionary most likely will not have a fast implementation (in contrast to the WP dictionary). And if the dictionary is learned on a large collection of music it might lead to a fairly large dictionary, since, even though the individual signals are sparse (in terms of time and frequency), the whole collection may not be, at least not sparse enough to overcome the lack of a fast implementation.

5 Acknowledgment

This work is supported by the Danish Technical Science Foundation (STVF), program no. 56-00-0143 (WAVES).

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: Design of dictionaries for sparse representation. In *Proc. of SPARS05*, pages 9–12, 2005.

[2] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Comput.*, 20(1):33–61, 1998.

[3] I. Daubechies. Time-frequency localization operators: A geometric phase space approach. *IEEE Trans. Inform. Theory*, 34:605–612, 1988.

[4] I. Daubechies. Orthonormal bases of compactly supported wavelets. II. Variation on a theme. *SIAM J. Math. Anal.*, 24(2):499–519, march 1993.

[5] L. Daudet, M. Sandler, and B. Torresani. Audio representation on overcomplete sets. *Proceedings of 14th conf. on Digital Signal Processing.*, 2002.

[6] D. Donoho and M. Elad. Optimally sparse representations in general (non-orthogonal) dictionaries via ℓ^1 minimization. *Proc. National Academy of Sciences of USA*, 100(5):2197 – 2202, March 2003.

[7] L. Ø. Endelt and A. la Cour-Harbo. Wavelets for sparse representation of music. In *Proc. of WEDELMUSIC*, 2004.

[8] L. Ø. Endelt and A. la Cour-Harbo. Energy distribution of coeficients of redundant signal representations of music. In *Proc. of SPARS05*, pages 33–36, 2005.

[9] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Information Theory*, 50:1341–1344, June 2004.

[10] I. Gorodnitsky and B. Rao. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Trans. Sig. Proc.*, 45(3):600 – 616, March 1997.

[11] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. Preprint, 2003.

[12] R. Gribonval and M. Nielsen. Sparse decomposition in "incoherent" dictionaries. In *Proc. IEEE Intl. Conf. on Image Proc.*, 2003.

[13] E. Hernández and G. Weiss. *A first course on wavelets*. CRC Press, Boca Raton, FL, 1996. With a foreword by Yves Meyer.

[14] J. Herredsvela, K. Engan, T. O. Gulsrud, and K. Skretting. Texture classification using sparse representations by learned compound dictionaries. In *Proc. of SPARS05*, pages 13–16, 2005.

[15] http://www.control.aau.dk/∼oertoft.

[16] S. Jaggi, W. Karl, S. Mallat, and A. Willsky. High resolution pursuit for feature extraction. *Applied and Computational Harmonic Analysis*, 5:428, October 1998.

[17] A. Jensen and A. la Cour-Harbo. *Ripples in Mathematics: The Discrete Wavelet Transform*. Springer, 2001.

[18] A. la Cour-Harbo. Application of the minimum fuel network to music signals. In *Proc. of IEEE Int. Conf. on Acou., Speech, and Sig. Proc.*, pages 301 – 304, 2004.

[19] S. Z. Li. Content-based audio classification and retrieval using the nearest feature line method. *IEEE Transactions on Speech and Audio Processing*, 8(5):619–625, September 2000.

[20] C.-C. Lin, S.-H. Chen, T.-K. Truong, and Y. Chang. Audio classification and categorization based on wavelets and support vector machine. *IEEE Transactions on Speech and Audio Processing*, 13(5), September 2005.

[21] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. on Sig. Proc.*, 41(12):3397 – 3415, 1993.

[22] E. Scheirer. Tempo and beat analysis of acoustic musical signals. *Journal of Acoustical Society of America*, pages 419–429, January 1998.

[23] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.

[24] Z. Wang, J. Cheung, Y. Xia, and J. Chen. Minimum fuel neural network and their applications to overcomplete signal representation. *IEEE Trans. Circuit and Systems*, 47(8):1146–1159, August 2000.

[25] M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. A K Peters, 1994.